UNLOCKING FROZEN POTENTIAL: DYNAMIC ANALYTICAL PROTOTYPES FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-Shot Class Incremental Learning challenges models to continuously learn novel categories from very few samples while mitigating the stability-plasticity dilemma. Recent methods increasingly leverage the robust generalization of Pre-Trained Models (PTMs) and stable prototype-based classifiers. We identify a critical limitation: **Rich Features but Constrained Potential**. Although PTMs provide rich generic representations, their potential is constrained by the frozen backbone's insufficient adaptation to novel concepts, the statistical bias of prototypes derived from very few samples and the absence of global discriminative optimization in static prototypes. Consequently, static prototype-based methods often struggle with feature entanglement and unreliable prototypes. To unlock the frozen potential, we propose **Dynamic Analytical Prototypes** that refine the post-backbone discriminative space. The first module Prior-Guided Distribution Calibration effectively improves the discriminative capability of representations by leveraging the base-class priors. Second, to achieve a global optimal decision, we derive a Global Knowledge-Aware Solver. This recursive solution updates class-balanced weights and theoretically mitigates the risk of catastrophic forgetting. Extensive experiments demonstrate consistent improvements over representative feature- and classifier-level optimization methods, validating its balance between stability and plasticity.

1 INTRODUCTION

Class-incremental learning (CIL) aims to acquire new semantic concepts in successive sessions while preserving previously learned knowledge. Conventional CIL can rely on substantial data for each new class and often uses distillation or rehearsal when the representation changes (Li & Hoiem, 2016; Rebuffi et al., 2017). Few-shot class-incremental learning (FSCIL) begins with abundant labeled data in a base session, then introduces disjoint novel classes from only a few examples without revisiting historical raw images (Tao et al., 2020). At every stage, predictions compete over all observed classes. This setting sharpens the stability–plasticity problem: the model must learn from scarce novel data while the abundant base data and limited access to past examples constrain subsequent classifier updates.

Early FSCIL methods mainly jointly trained a feature extractor and a classifier from scratch, but extreme data scarcity often caused novel-class overfitting and severe forgetting. Recent approaches therefore exploit the generic representations of Pre-Trained Models (PTMs) and stable prototype classifiers such as Nearest Class Mean (NCM)Guerriero et al. (2018). This combination provides rich features and ensures stability. Recent PTM-based methods largely adopt the Parameter-Efficient Fine-Tuning (PEFT) paradigm, which freezes the backbone to further ensure stability while introducing learnable modules (e.g., promptsJia et al. (2022) and adaptersChen et al. (2022b)) to transfer the rich priors to downstream tasks. However, a critical limitation is exposed in the current state-of-the-art: **Rich Features but Constrained Potential**.

To unlock more of this decision potential, we propose Dynamic Analytical Prototypes (DAP) with two complementary components. Prior-guided distribution calibration (PGDC) uses base-class statistics to define surrogate novel-class distributions, samples features from them, and places base

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

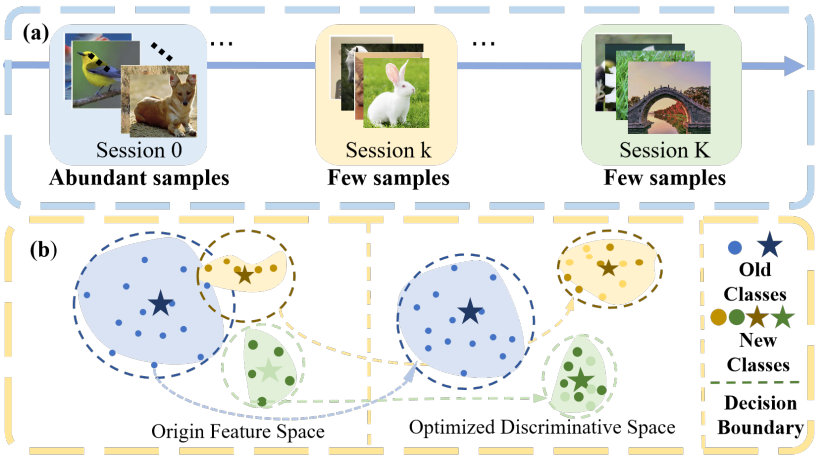

Figure 1: The FSCIL setting and DAP's two classifier-side opportunities: (a) successive base and few-shot sessions; (b) calibrating uncertain novel statistics and updating decisions over competing classes.

and novel features in a shared space of fixed Gaussian-RBF anchors. The global knowledge-aware solver (GKAS) then turns the mapped class summaries into decision prototypes through a class-balanced analytical objective. As new classes arrive, the shared state updates both old and new decision weights without accessing historical raw images. The backbone is trained or adapted on the base session and remains frozen in subsequent sessions.

Our contributions follow the original two-stage method. First, we identify the classifier-side gap between rich frozen features and uncertain, static few-shot prototypes. Second, We propose Dynamic Analytical Prototypes, a novel stable and dynamic paradigm that unlocks the frozen potential of PTMs. This paradigm dynamically rectifies and updates the biased prototypes, achieving global optimal decision on discriminative features without catastrophic forgetting. Third, We conduct controlled evaluations across standard FSCIL benchmarks and multiple frozen backbones. DAP consistently improves representative post-backbone baselines and can be integrated with existing FSCIL frameworks, confirming that our framework successfully establishes an equilibrium between plasticity and stability.

2 Related Work

2.1 Class-Incremental Learning

Early CIL methods addressed forgetting while learning or adapting a backbone. Learning without Forgetting and iCaRL preserve earlier responses through distillation; Riemannian Walk constrains parameter changes, and gradient-based sample selection retains examples for rehearsal (Li & Hoiem, 2016; Rebuffi et al., 2017; Chaudhry et al., 2018; Aljundi et al., 2019). ACIL Zhuang et al. (2022) first introduced recursive closed-form learning into CIL. However, its conventional sample-wise objective is not designed for FSCIL, where unreliable few-shot statistics and extreme base–novel imbalance can skew the analytical solution toward base classes.

With pretrained models, one direction adapts a backbone using prompts or adapters, as in L2P, DualPrompt, and CODA-Prompt (Jia et al., 2022; Chen et al., 2022a; Wang et al., 2022b;a; Smith et al., 2023). Another exploits fixed features: RanPAC combines nonlinear random projection with an analytical head, whereas FeCAM uses class-dependent covariance-aware distances (McDonnell et al., 2023b; Goswami et al., 2023). Across frameworks, the Nearest Class Mean (NCM) classifier remains the dominant choice due to its non-parametric nature and representative prototypes derived from abundant samples, serving as the classification template during inference. However, the static prototype-based classifier is often a biased estimator of the true class distribution due to the insufficient adaptation to novel concepts and severe data scarcity in the FSCIL regime. This limitation underscores the critical need for robust and dynamic solver designs tailored for FSCIL.

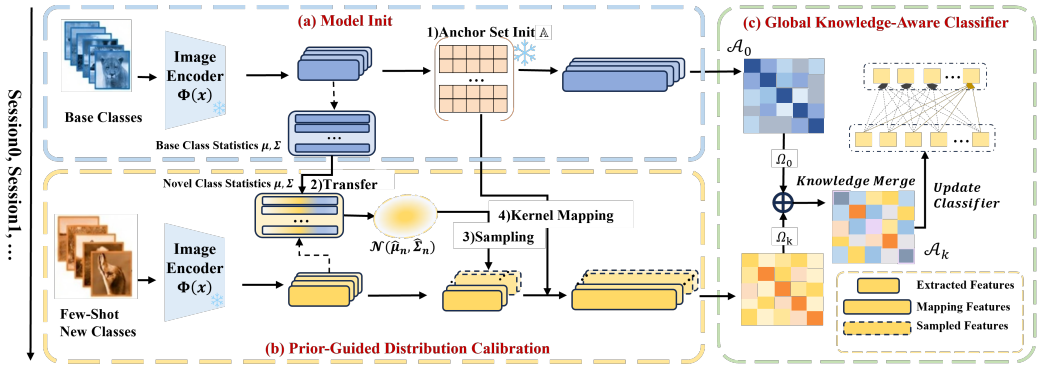

Figure 2: DAP pipeline: base-session adaptation establishes features and fixed Gaussian anchors; PGDC transfers base statistics to a novel-class surrogate distribution; GKAS recursively updates analytical decision prototypes. The representation stays fixed during incremental sessions.

2.2 Few-Shot Class-Incremental Learning

Early FSCIL work sought to balance new-class learning and old-class stability through replay, subspace regularization, or representation-learning objectives (Tao et al., 2020; Liu et al., 2022; Akyürek et al., 2021; Ji et al., 2023). GKEALZhuang et al. (2023) first extends analytical learning to FSCIL by employing Augmented Feature Concatenation to increase the contribution of novel classes. However, since the augmented features are still derived from the same statistically unreliable few-shot support, this strategy cannot explicitly correct the underlying distribution estimation error and may even amplify its influence on the analytical solution, leading to suboptimal novel-class performance.

Pretrained FSCIL methods further use prompts, visual–language supervision, or approximation-aware training, as in ASP, PriViLege, and ApproxFSCIL (Liu et al., 2024; Park et al., 2024; Wang et al., 2024). Recognizing that prototypes derived from few samples are inherently biased, methods like TEENWang et al. (2023) and BiMCChen et al. (2025) design prototype rectification method to estimate more representative class centers. Despite these advancements, most existing methods freeze their prototype-based classifier(e.g., NCM) to ensure stability. However, within this dynamic data stream, static NCM lacks the global discriminative context. Inspired by Goswami et al. (2024), we explicitly rectify the statistical bias of few-shot prototypes from the prior of base classes. Moreover, we endow the static prototypes with deriving capability, achieving a class-balanced and theoretical optimal update that NCM and ACIL inherently lacks.

3 Method

3.1 Preliminaries

Problem setup. In FSCIL, the model observes sessions $D_0, D_1, \ldots, D_T$ with disjoint label sets $\mathcal{C}_t$. The base session has abundant training examples, whereas each subsequent class has five labeled examples in the reported benchmarks. Evaluation after session t includes every class in $\mathcal{C}_{0:t} = \bigcup_{s=0}^{t} \mathcal{C}_s$, with all scores in one competition. We adapt the pretrained feature extractor $\phi : \mathcal{X} \to \mathbb{R}^d$ on D_0 and freeze it for $t > 0$. Hence the incremental classifier update needs no backbone gradient training, although base adaptation is a training step.

Model formulation. The FSCIL model can be generally decomposed into a feature extractor and classification heads**?**, i.e., $f(x) = \mathcal{W}^\top \phi(x)$. The empirical class prototype is

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} \phi(x_i). \tag{1}$$

3.2 PRIOR-GUIDED DISTRIBUTION CALIBRATION

Although some methods have proposed to explicitly calibrate the biased prototype for new classes, they neglect the insufficient adaptation to novel concepts of PTMs, leading to the feature entanglement among semantically similar categories. To address these limitations, drawing inspiration from Goswami et al. (2024), we leverage the statistically reliable distribution of base classes as prior knowledge to dynamically rectify the distorted novel distributions and disentangle the rectified features, enhancing the discrimination capability of these features.

For class c, let Σ_c be its empirical feature covariance and $\mathcal{B} = \mathcal{C}_0$ the base classes. The cosine affinity between a novel prototype μ_n and a base prototype μ_b determines a soft transfer weight:

$$S_{n,b} = \frac{\mu_n^\top \mu_b}{\|\mu_n\|\|\mu_b\|}, \qquad w_{n,b} = \frac{\exp(S_{n,b}/\tau)}{\sum_{j\in\mathcal{B}} \exp(S_{n,j}/\tau)}. \tag{2}$$

The calibrated mean and covariance, followed by the reported synthetic-feature draw, are

$$\hat{\mu}_n = \alpha\mu_n + (1-\alpha)\sum_{b\in\mathcal{B}} w_{n,b}\mu_b, \tag{3}$$

$$\hat{\Sigma}_n = \beta\left(\Sigma_n + \sum_{b\in\mathcal{B}} w_{n,b}\Sigma_b\right), \tag{4}$$

$$\hat{\phi}(x) \sim \mathcal{N}(\hat{\mu}_n, \hat{\Sigma}_n). \tag{5}$$

Here $\tau > 0$ controls the concentration of base-class weights, $\alpha \in [0,1]$ retains part of the novel support mean, and $\beta > 0$ scales the synthetic covariance.

Though rectified, this statistical procedure alone does not decouple the entangled feature manifold. To resolve this, we employ a distance-based kernel mapping to construct a high-dimensional reference space induced by the rich and diverse representations learned from the base session. This reference space provides a structural scaffold for encoding the global semantic geometry. Specifically, we independently and identically sample a subset of features from the base-session dataset $\mathcal{D}_0$ as kernel anchors, i.e., $\mathbb{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$, where each $\alpha_k \in \mathbb{R}^d$ is a feature vector (typically $K \gg d$). This anchor set remains fixed throughout the incremental learning process to ensure a consistent kernel space.

For any incoming feature vector $\phi(x)$ (belonging to the base session $\mathcal{D}_0$ or a novel few-shot session $\mathcal{D}_t$), we construct its high-dimensional embedding $\Psi(x) \in \mathbb{R}^K$ by measuring its kernel distance to the pre-defined anchors. This transformation quantifies the relative position of input features within the global space. The mapping function $\Psi : \mathbb{R}^d \to \mathbb{R}^K$ is formulated as:

$$\Psi(x) = [\kappa(\phi(x), \alpha_1), \kappa(\phi(x), \alpha_2), \ldots, \kappa(\phi(x), \alpha_K)]^\top \tag{6}$$

where the kernel function $\kappa(\cdot,\cdot)$ is defined as the Gaussian Radial Basis Function (RBF):

$$\kappa(\phi(x), \alpha_k) = \exp\left(-\gamma\|\phi(x) - \alpha_k\|^2\right) = \exp(-\gamma(\|\phi(x)\|^2 + \|\alpha_k\|^2) \cdot \exp(2\gamma\langle\phi(x), \alpha_k\rangle) \tag{7}$$

Here, γ is the bandwidth hyperparameter controlling the receptive field of each anchor. By projecting data onto this K-dimensional kernel space anchored on historical knowledge, the entangled manifold in the original space prone to be decoupled in the new feature space $\Psi(x)$. We provide theoretical propositions to prove why this specific mapping is effective (detailed proof is provided in the supplementary materials):

Cover's Theorem. Cover (1965) The theoretical foundation for high dimensional mapping is to break the linear inseparability of the frozen feature space. For a set of N patterns in a L-dimensional space, the probability that a random dichotomy is linearly separable is negligible when $N \gg L$. However, by mapping these patterns non-linearly into a higher-dimensional space $\mathbb{R}^K$ where $K \geq N$, the probability of linear separability approaches 1.

Proposition 3.1. High-order interaction. Beyond dimension elevation, the Gaussian kernel implicitly introduces non-linear complexity essential for distinguishing fine-grained differences. Applying Taylor expansion to the interaction term $\exp(\gamma\langle\phi(x), \alpha\rangle)$, it implicitly encapsulates all n-th

order cross-products between feature dimensions:

$$\exp\left(\gamma\langle\phi(x),\alpha\rangle\right)=\sum_{n=0}^{\infty}\frac{(2\gamma)^n}{n!}\langle\phi(x)^\top\alpha\rangle^n \tag{8}$$

Proposition 3.2. Stability and Separability. We prove that large K induces geometry stability. Applying Hoeffding's inequality and Cauchy-Schwarz Inequality, we make 2 observations. As K increases in Eq.(9), the feature norm concentrates around its expectation with exponential stability, ensuring features reside on stable hyperspheres regardless of sampling bias.

$$\mathbb{P}\left(\left|\frac{1}{K}\|\Psi(x)\|^2-\mu_{self}\right|\geq\epsilon\right)\leq 2\exp(-2K\epsilon^2) \tag{9}$$

where $\mu_{self}=\mathbb{E}(\frac{1}{K}\|\Psi(x)\|^2)$ as the expected-energy. Furthermore, as K increases, the expected cross-correlation is strictly bounded by their respective self-correlations, as formulated in Eq.(10).

$$\mathbb{E}[\kappa(x,\alpha)\cdot\kappa(x',\alpha)]<\sqrt{\mathbb{E}[\mu_{self}(x)]\mathbb{E}[\mu_{self}(x')]} \tag{10}$$

The strict inequality in this formulation is crucial. It explicitly indicate that the mutual affinity between representations from different classes (e.g., between new and old classes) remains strictly inferior to the mean of their internal intra-class correlations. This definitive energy gap forces intra-class compactness to consistently dominate over inter-class interference, thereby providing a theoretical guarantee for strict linear separability in the high-dimensional kernel space.

Overall, this joint calibration and mapping process not only effectively reduces the statistical bias derived from few samples, but reshapes the representation manifold, constructing a well-separated feature space that enables robust classification even with a linear solution.

3.3 Global Knowledge-Aware Solver

Although the high-dimensional kernel space theoretically guarantees separability, it inevitably introduces feature sparsity and noise. Static NCM is insufficient here, as it implicitly focuses on the intra-class compactness but lacks the global view to penalize confusion with similar categories, thereby constraining the discriminative potential of these calibrated features. Meanwhile, conventional analytic methods such as ACIL adopt an unweighted sample-wise least-squares objective, which does not explicitly account for the extreme base–novel imbalance in FSCIL. To unlock this constraint, we propose a global knowledge-aware solver. It recursively updates all prototypes with every new session and is mathematically equivalent to retraining on all data.

Let $\Phi_{0:t}\in\mathbb{R}^{N_{0:t}\times K}$ denote the matrix of kernel-mapped features accumulated up to session t, and $\mathcal{Y}_{0:t}\in\mathbb{R}^{N_{0:t}\times C}$ be the corresponding one-hot label matrix. Our goal is to learn a projection matrix $\mathcal{W}_t\in\mathbb{R}^{K\times C}$ that minimizes the regularized squared error. Crucially, to prevent preference of specific class due to the extreme imbalance between abundant base classes and scarce novel classes, we derive a class-adaptive strategy. It gives each sample a weight equal to inversely proportional to its class frequency. The objective function is formulated as:

$$\mathcal{L}_{\mathcal{W}_t}=\|\Omega_{0:t}^{\frac{1}{2}}(\mathcal{Y}_{0:t}-\Phi_{0:t}\mathcal{W}_t)\|_F^2+\lambda\|\mathcal{W}_t\|_F^2 \tag{11}$$

where $\lambda>0$ is a regularization coefficient that ensures numerical stability and $\Omega_{0:t}\in\mathbb{R}^{N_{0:t}\times N_{0:t}}$ is a diagonal weight matrix. Furthermore, the optimization problem in Eq.(11) constitutes a strictly convex least squares problem, it admits an analytical solution. By applying the method of least squares (i.e., setting the derivative w.r.t. $\mathcal{W}$ to zero) , we obtain the initial project matrix at session 0:

$$\mathcal{W}_0^*=\left(\Phi_0^\top\Omega_0\Phi_0+\lambda I\right)^{-1}\Phi_0^\top\Omega_0\mathcal{Y}_0 \tag{12}$$

Here, I is an identity matrix. $\Phi_0^\top\Omega_0\mathcal{Y}_0$ is mathematically equivalent to the empirical prototypes, and $\mathcal{A}_0=\Phi_0^\top\Omega_0\Phi_0+\lambda I$ denotes the global knowledge gram matrix. By multiplying with $\mathcal{A}_0^{-1}$, the prototypes explicitly suppress the interference among correlated categories, rectifying the local empirical means into globally optimized decision anchors(Proof see Supplementary Material). Consequently, we term the resulting projection matrix $\mathcal{W}$ as Dynamic Analytical Prototypes.

To fit the dynamic data stream, i.e., for session $t+1$ (where $t \geq 0$), where access to historical data $\Phi_{0:t}$ is prohibited, we observe that the prototype matrix $\mathcal{W}$ satisfies the recursive property. We can decompose it into the historical state and the current incremental update via the following proposition. Firstly, we decompose the positive-definite gram matrix $\mathcal{A}$ via Cholesky-decomposition:

$$\mathcal{A}_t^{-1} = \mathcal{P}_t \mathcal{P}_t^\top \tag{13}$$

Proposition 3.3 Recursive Update Solver. (Detailed proof is provided in the supplementary materials.) Given calibrated $\mathcal{D}_{t+1}^{train}$, global knowledge matrix $\mathcal{A}_k^{-1}$ and historical prototypes $\mathcal{W}_k$, we decompose Eq.(11) into multi tasks and obtain $\mathcal{W}_{k+1}$ recursively as:

$$\mathcal{W}_{t+1} = \left[\mathcal{W}_t - \mathcal{A}_{t+1}^{-1} \Phi_{t+1}^\top \Omega_{t+1} \Phi_{t+1} \mathcal{W}_t \;\; \mathcal{A}_{t+1}^{-1} \mu_{t+1}^\top \right] \tag{14}$$

where

$$\begin{aligned} \mathcal{A}_{t+1}^{-1} &= \mathcal{A}_t^{-1} - \mathcal{A}_t^{-1} \Phi_{t+1}^\top (\Omega_{t+1}^{-1} + \Phi_{t+1}^T \mathcal{A}_t^{-1} \Phi_{t+1})^{-1} \Phi_{t+1} \mathcal{A}_t^{-1} \\ &= \mathcal{P}_t (I - \mathcal{Q}_t^\top (\Omega_{t+1}^{-1} + \mathcal{Q}_t \mathcal{Q}_t^\top)^{-1} \mathcal{Q}_t) \mathcal{P}_t^\top \end{aligned} \tag{15}$$

and

$$\mu_{t+1}^\top = \Phi_{t+1}^\top \Omega_{t+1} \mathcal{Y}_{k+1} \quad \mathcal{Q}_t = \Phi_{t+1} \mathcal{P}_t \tag{16}$$

Proposition 3.3 dynamically integrates historical and novel prototypes via the global knowledge matrix $\mathcal{A}_{t+1}$. By balancing accumulated knowledge with incoming statistics, the class-balanced weights prevent abundant base classes from dominating the analytical objective, thereby alleviating the extreme base–novel imbalance in FSCIL. Meanwhile, the recursive injection reshapes the decision boundaries as if it had accessed the entire historical dataset, yielding a globally optimal solution. Through the tight coupling of PGDC and GKAS, our framework transforms unreliable 5-shot novel statistics into global old/new decisions under frozen features.

4 Experiments

4.1 Experimental Setup

Dataset. In line with Zhou et al. (2025), we evaluate our method on three standard FSCIL benchmarks: ImageNetRHendrycks et al. (2021), CUB200Wah et al. (2011) and CIFAR100Krizhevsky et al. (2009). CIFAR100 is divided into 60 base classes and 40 new classes. where the new classes are introduced through eight 5-way 5-shot incremental sessions. CUB200 and ImageNetR each contain 100 base and 100 novel classes, organized into ten 10-way 5-shot incremental sessions.

Evaluation Metrics. In line with Zhou et al. (2025), we report the Top-1 accuracy on the accumulated test set $\mathcal{D}_{0:k}^{test}$. We also utilize the average accuracy(AVG) between all sessions to evaluate the entire model, and the performance drop(PD) between the base and final session accuracies to quantify catastrophic forgetting. For ablation studies and parameter sensitivity analysis, we primarily report the AVG to offer a holistic performance assessment.

Implementation Details. To evaluate the architecture-agnostic nature of our method, we employ several pretrained models as feature extractors. Crucially, the backbone is strictly frozen across all subsequent incremental sessions to ensure representational stability. The hyperparameters α, β, K, γ, τ are empirically set to {0.6, 0.8, 5000 , 1.0, 16}for CIFAR100, {0.8, 0.8, 3000, 0.1, 25} for CUB200 and { 0.6, 0.7, 5000, 1.0,50} for ImageNetR. To ensure a fair benchmarking against representation methods, our main comparisons are all conducted using the pretrained ResNet-18, VIT and CLIP-VIT-B architectures to evaluate cross-architecture generalizability. The supplementary material provides the complete algorithmic pipeline and detailed training configurations, together with full results on ImageNetR and CIFAR100, session-wise accuracy curves, comprehensive ablation studies, t-SNE visualizations, and failure-case analysis.

Method	Backbone	CIFAR-100			CUB-200			ImageNet-R		
		A_{base}	A_{last}	AVG	A_{base}	A_{last}	AVG	A_{base}	A_{last}	AVG
NCM	ResNet-18	78.72	60.16	68.07	74.65	55.66	63.18	68.90	45.89	54.84
TEEN		78.72	61.44	68.88	74.65	55.42	62.55	68.90	47.04	55.55
GKEAL		83.53	51.18	65.21	**77.09**	57.68	65.02	70.89	48.41	57.27
C-FeCAM		80.35	61.17	69.86	75.98	57.06	64.78	68.97	47.55	56.03
C-RanPAC		82.73	61.15	70.28	77.03	57.42	64.72	70.21	47.65	56.48
DAP		**83.60**	**64.82**	**72.92**	**77.09**	**58.25**	**65.29**	**71.38**	**49.29**	**58.20**
NCM	VIT-B-IN21K	92.02	85.40	88.01	87.57	80.98	82.63	82.68	69.62	74.86
TEEN		92.02	86.06	88.44	87.57	78.89	81.57	82.68	70.36	75.58
GKEAL		**94.00**	87.51	89.93	89.12	78.40	82.18	86.91	71.11	78.26
C-FeCAM		92.93	74.77	81.72	87.60	77.42	80.49	84.60	58.30	68.27
C-RanPAC		93.78	85.94	88.99	89.00	81.33	84.37	86.39	75.37	80.21
DAP		**94.00**	**88.21**	**90.61**	**90.29**	**84.07**	**85.89**	**87.20**	**75.70**	**80.75**
NCM	CLIP-VIT-B	85.50	67.92	75.37	82.26	67.62	72.89	82.61	63.38	70.98
TEEN		85.50	68.57	75.74	82.26	65.76	71.75	82.61	63.55	71.38
GKEAL		86.87	69.85	76.65	82.44	68.05	73.42	81.14	64.32	72.99
C-FeCAM		85.62	68.94	76.02	82.23	68.19	73.34	82.58	64.75	71.95
C-RanPAC		86.78	66.60	74.78	82.72	68.55	73.64	83.72	65.93	72.68
DAP		**87.02**	**71.79**	**78.28**	**83.62**	**69.30**	**74.52**	**84.01**	**67.18**	**73.98**
SEC-Prompt	VIT-B-IN1K	91.23	86.12	88.25	86.59	81.42	82.45	84.30	74.25	78.30
SEC-Prompt+DAP		93.08	89.38	91.02	90.26	**85.16**	**86.22**	**86.15**	**77.68**	**81.24**
CASP		93.07	87.04	89.33	88.41	82.91	83.92	82.20	70.95	75.31
CASP+DAP		**94.**23	**88.43**	**90.67**	**90.22**	**84.79**	**86.12**	**85.32**	**74.08**	**78.83**

Table 1: Source-reported top-1 accuracy (%). $A_{\text{base}} = A_0$ is accuracy at the end of base-session learning, not final-session base-class accuracy; $A_{\text{last}} = A_T$ tests all seen classes; AVG averages complete sessions. Bold values identify the maximum in each shared-backbone head-comparison block for each dataset and metric

4.2 COMPARISON

We compare DAP with representative frozen-feature optimization methods, including prototype calibration (TEEN), higher-order distribution calibration (C-FeCAM), random-projection-based classification (C-RanPAC), and few–shot analytical learning (GKEAL). To isolate the contribution of post-backbone optimization, all methods are evaluated with the same frozen feature extractor under three backbone families: ResNet-18, VIT-B-IN21K, and CLIP-VIT-B. We further integrate DAP into the PTM-based FSCIL frameworks CASP Huang et al. (2026) and SEC-Prompt Liu & Yang (2025) to evaluate its compatibility with existing representation-learning approaches. We report the average results over three random seeds in Table 1.

DAP demonstrates strong performance across different representation architectures. With VIT-B-IN21K, DAP achieves the highest average accuracy of 85.89%, outperforming the strongest baseline C-RanPAC by 1.52 percentage points. Under CLIP-ViT-B, DAP also obtains the best average accuracy of 74.52%, improving over C-RanPAC by 0.88 points. On ResNet-18, DAP remains competitive with an average accuracy of 65.29%. These results indicate that DAP can effectively improve class discrimination across CNN, supervised Transformer, and vision–language pretrained representations.Although DAP does not always achieve the lowest performance drop, it maintains the highest absolute accuracy across incremental sessions. This suggests that DAP achieves a stronger overall stability–plasticity trade-off by effectively acquiring novel classes while preserving competitive performance on previously learned knowledge.

We further integrate DAP into CASP and SEC-Prompt to examine its compatibility with existing FSCIL frameworks. DAP improves the average accuracy from 83.92% to 86.12% and 82.45% to 86.22%, while reducing their performance drop. These results show that DAP can serve as a complementary post-backbone optimization module for existing frameworks.

DC	KM	GKAS	CIFAR-100		CUB-200		ImageNet-R	
			AVG	PD	AVG	PD	AVG	PD
–	–	–	88.10	6.33	82.78	6.54	75.20	11.85
✓	–	–	88.76	**5.29**	82.75	6.68	75.91	10.81
–	✓	–	80.68	10.07	73.22	11.20	60.21	18.73
–	–	✓	88.75	7.45	83.59	6.91	77.95	11.40
✓	–	✓	89.85	5.94	83.59	7.15	78.86	**10.13**
✓	✓	–	79.90	11.42	72.50	12.29	60.00	19.22
–	✓	✓	89.61	7.18	85.53	6.57	80.08	12.44
✓	✓	✓	**90.61**	5.79	**85.89**	**6.22**	**80.75**	11.50

Table 3: Component ablation on the base-adapted ViT-B-IN21K backbone (%). Bold indicates the highest AVG or lowest PD within each dataset.

Source-reported update	AVG (%)	Time (ms)	Memory (MB)
Batch recomputation	85.89	530.21	1229.77
Recursive GKAS	85.89	92.63	207.61

Table 2: Reported CUB-200 batch and recursive efficiency.

4.3 Ablation Study

To investigate the individual contribution of each component in our proposed framework, we conducted ablation studies on all the datasets and mainly analyzed the result of CUB200. We further separate DAP into three parts, i.e., Distribution Calibration (DC), Kernel Method (KM) and Global Knowledge-Aware Solver (GKAS). The quantitative results are summarized in Table 3. The first row acts as the baseline, i.e., the VIT-B-IN21K and NCM Classifier.

Effectiveness and necessity of DC The primary function of the DC module is to rectify the statistical bias inherent in few-shot novel classes by transferring priors from abundant base classes. Removing DC from the full framework decreases AVG from 85.89% to 85.53% and increases PD from 6.22% to 6.57%. These results show that kernel mapping and GKAS cannot fully compensate for the estimation errors caused by extreme data scarcity, making distribution calibration an important component of DAP.

Effectiveness and necessity of Kernel Mapping (KM). KM maps the calibrated features into a nonlinear high-dimensional reference space to improve their separability. Removing KM from the full framework decreases AVG from 85.89% to 83.59% and increases PD from 6.22% to 7.15%. This result indicates that distribution calibration alone cannot fully eliminate the residual overlap among similar classes, limiting the effectiveness of subsequent global optimization. KM therefore serves as a critical bridge between DC and GKAS by transforming the calibrated class distributions into a more discriminative geometry for analytical classification.

Effectiveness and necessity of the Global Knowledge-Aware Solver (GKAS). GKAS performs class-balanced global optimization over the mapped features. With the support of DC and KM, GKAS achieves the best AVG of 85.89% and the lowest PD of 6.22%. Conversely, removing GKAS from PGDC reduces AVG to 72.50%, since static NCM cannot effectively exploit the sparse and correlated high-dimensional representations. GKAS thus converts the calibrated and remapped features into global, class-balanced decision boundaries.

DAP therefore tightly couples the modules: PGDC improves few-shot novel statistics, Gaussian anchors align old and novel classes in a shared geometry, and GKAS converts this calibrated kernel space into a class-balanced global decision objective. Therefore, the modules are mutually compensatory and removing any of them breaks the connection between statistical reliability, feature separability, and global discrimination.

Furthermore, we compare our recursive GKAS against a non-recursive baseline (recomputing with all accumulated historical data) under strict double-precision to guarantee numerical stability. As

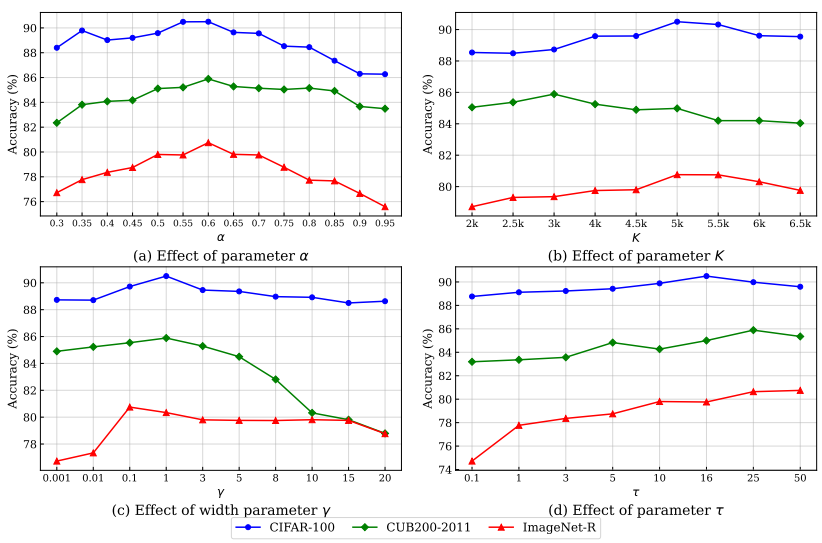

Figure 3: Performance on various dataset of a) calibration parameter α, b) width parameter γ, c) Kernel Size K and d) temperature parameter τ

illustrated in Table 2 , they Both obtain 85.89% AVG. while recursion reduces update time from 530.21 to 92.63 ms and peak memory from 1229.77 to 207.61 MB. This strictly validates the superior computational efficiency of our design.

Hyper-Parameter Analysis We further examine the sensitivity of the key PGDC hyperparameters, including α, β, τ, K, and γ. As shown in Figure 3, DAP remains stable across a broad range of settings.

The calibration coefficients α and β control the balance between the original novel-class statistics and the transferred base-class priors. The model is largely insensitive to β, allowing it to be fixed across experiments. In contrast, an overly small α causes excessive reliance on base-class priors and may homogenize novel-class representations, whereas an overly large α weakens the benefit of distribution calibration.

The temperature τ determines the concentration of the base-class transfer weights. A very small τ restricts calibration to only a few neighboring classes, while a large τ may introduce statistics from irrelevant classes. The results generally favor a relatively concentrated weighting distribution with $\tau \geq 10$.

The kernel dimension K controls the capacity of the mapped feature space. Increasing K initially improves discrimination, but excessively large values introduce additional noise, memory consumption, and computational cost. Therefore, an intermediate K provides the best trade-off between feature separability and efficiency.

Finally, the bandwidth γ determines the receptive range of the RBF kernel. Excessively large values produce sparse, nearly orthogonal embeddings and degrade classification performance.

5 Conclusion

We proposed Dynamic Analytical Prototypes (DAP) to better exploit frozen representations for FSCIL. PGDC calibrates novel-class statistics and constructs a shared kernel space, while GKAS recursively updates class-balanced decision prototypes without revisiting historical images. Results on three benchmarks show higher average accuracy than representative baselines across multiple backbones, supporting the integration of distribution calibration and analytical prototype updates.

REFERENCES

Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059*, 2021.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–547, 2018.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022a.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022b.

Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, and Wenbin Li. Enhancing few-shot class-incremental learning via training-free bi-level modality calibration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9881–9890, 2025.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost Van De Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36:6582–6595, 2023.

Dipam Goswami, Bartłomiej Twardowski, and Joost van de Weijer. Calibrating higher-order statistics for few-shot class-incremental learning with pre-trained vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4075–4084, 2024. URL https://arxiv.org/abs/2404.06622.

Samantha Guerriero, Barbara Caputo, and Thomas Mensink. Deepncm: Deep nearest class mean classifiers. 2018.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Shuai Huang, Xuhan Lin, and Yuwu Lu. Casp: Few-shot class-incremental learning with cls token attention steering prompts. *Pattern Recognition*, 177:113306, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113306. URL https://www.sciencedirect.com/science/article/pii/S0031320326002712.

Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Xuelong Li. Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing*, 32: 937–948, 2023.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629, 2016. doi: 10.1007/978-3-319-46493-0_37.

Chenxi Liu, Zhenyi Wang, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Few-shot class incremental learning with attention-aware self-adaptive prompt. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20053-3_9.

Ye Liu and Meng Yang. Sec-prompt: Semantic complementary prompting for few-shot class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25643–25656, 2025.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023a.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2793dc35e14003dd367684d93d236847-Paper-Conference.pdf.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23881–23890, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Rebuffi_iCaRL_Incremental_Classifier_CVPR_2017_paper.html.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11909–11919, 2023.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, and Yihong Gong. Few-shot class-incremental learning. *arXiv preprint arXiv:2004.10956*, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. *Advances in Neural Information Processing Systems*, 36:15060–15076, 2023.

Xuan Wang, Zhong Ji, Xiyao Liu, Yanwei Pang, and Jungong Han. On the approximation risk of few-shot class-incremental learning. In *European Conference on Computer Vision*, pp. 162–178. Springer, 2024.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. Acil: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022.

Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. Gkeal: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7746–7755, 2023.

SUPPLEMENTARY MATERIAL FOR UNLOCKING FROZEN POTENTIAL: DYNAMIC ANALYTICAL PROTOTYPES FOR FEW-SHOT CLASS-INCREMENTAL LEARNING

A OVERALL ALGORITHM

Algorithm 1 summarizes the complete learning procedure of DAP. The backbone is first adapted using only the base session and is then frozen throughout incremental learning. Base-class statistics and a fixed anchor set are retained to define a consistent kernel reference space. At each incremental session, DAP calibrates the unreliable novel-class statistics, maps the resulting features into the shared kernel space, and recursively updates a class-balanced analytical classifier using the accumulated analytical state. During inference, all query samples are processed by the frozen backbone and the fixed kernel mapping before classification with the current analytical weights.

B DETAILED PROOF OF GAUSSIAN KERNEL MAPPING

In this section, we provide rigorous mathematical proofs for the theorems presented in the main paper, validating the mechanism of our Gaussian Kernel Mapping in terms of high-order feature interactions, geometric stability, and linear separability.

B.1 THEORETICAL FOUNDATION: COVER'S THEOREM

Theorem (Cover's Theorem on Separability of Patterns): For a set of N patterns in an L-dimensional space, the probability that a random dichotomy is linearly separable is negligible when $N \gg L$. However, by mapping these patterns non-linearly into a higher-dimensional space $\mathbb{R}^K$ where $K \geq N$, the probability of linear separability approaches 1.

B.2 PROOF OF PROPOSITION 3.1:IMPLICIT HIGH-ORDER INTERACTIONS

Proposition 3.1. The Gaussian kernel mapping implicitly introduces non-linear complexity by encapsulating infinite-order feature interactions. The mapped feature space contains all n-th order cross-products of the original feature dimensions.

Proof. Let $x, \alpha \in \mathbb{R}^d$ denote the input feature vector and a kernel anchor, respectively. The Gaussian RBF kernel is defined as:

$$\kappa(x, \alpha) = \exp(-\gamma \|x - \alpha\|^2) \tag{17}$$

Expanding the squared Euclidean norm $\|x - \alpha\|^2 = \|x\|^2 + \|\alpha\|^2 - 2\langle x, \alpha \rangle$, we can factorize the kernel as:

$$\kappa(x, \alpha) = \exp(-\gamma \|x\|^2) \cdot \exp(-\gamma \|\alpha\|^2) \cdot \exp(2\gamma \langle x, \alpha \rangle) \tag{18}$$

Here, the first two terms depend solely on the norms of x and α, acting as instance-specific scaling factors. The core interaction between the query sample and the anchor lies in the term

Algorithm 1 Overall Learning Procedure of DAP

Require: Session stream $\{\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_T\}$; pretrained backbone ϕ; anchor number K; synthetic sample number N_{syn}; hyperparameters $\alpha, \beta, \tau, \gamma, \lambda$
Ensure: Analytical classifier $\mathbf{W}_t$ at each session
Adapt ϕ using only the base session $\mathcal{D}_0$, and then freeze the backbone
Extract base features and compute the mean and covariance of each base class
Sample K base features as the fixed anchor set $\mathcal{A} = \{\alpha_1, \ldots, \alpha_K\}$
Map the base features into the kernel space
Construct the class-balanced weight matrix $\mathbf{\Omega}_0$
Initialize $\mathbf{A}_0^{-1}$ and $\mathbf{W}_0$
for $t = 1$ **to** T **do**
  Extract the support features of session $\mathcal{D}_t$
  for each novel class c in session t **do**
    Estimate its empirical mean and covariance
    Compute base-class transfer weights using prototype similarity
    Calibrate the novel-class distribution
    Sample N_{syn} features from the calibrated distribution
  end for
  Map the calibrated features into the fixed kernel space according to Eq. (17)
  Construct the session-wise class-balanced matrix $\mathbf{\Omega}_t$
  Recursively update $\mathbf{A}_t^{-1}$ and $\mathbf{W}_t$ according to Eq. (38)
end for
Predict each query using the frozen backbone, fixed kernel mapping, and current classifier $\mathbf{W}_t$
return $\mathbf{W}_t$

$\exp(2\gamma\langle x, \alpha\rangle)$.To reveal the implicit feature interactions, we apply the Taylor Series expansion to this interaction term. Recall that $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$. Substituting $z = 2\gamma\langle x, \alpha\rangle$:

$$\exp(2\gamma\langle x, \alpha\rangle) = \sum_{n=0}^{\infty} \frac{(2\gamma)^n}{n!} (\langle x, \alpha\rangle)^n \tag{19}$$

Conclusion: This proves that our mapping implicitly projects the original d-dimensional features into an infinite-dimensional space spanned by all possible n-th order interactions (cross-products) between feature dimensions. This high-order information is crucial for capturing fine-grained semantic differences that are linearly inseparable in the original low-order space.

B.3 Proof of Proposition 3.2: Stability and Separability

Proposition 3.2. With a sufficiently large number of anchors K, the mapped features satisfy two key properties: Stability: The feature norm concentrates around its expectation with exponential probability. Separability: Distinct classes exhibit a strict energy gap between intra-class compactness and inter-class correlation, ensuring linear separability.

Proof of Geometric Stability (Concentration of Measure) The kernel anchors $\mathbb{A} = \{\alpha_1, \ldots, \alpha_K\}$ are sampled independently and identically (i.i.d.) from the base data distribution $\mathcal{D}_0$.

Proof. Consider the normalized squared norm (energy) of the mapped feature $\Psi(x) \in \mathbb{R}^K$:

$$Z_K = \frac{1}{K} \|\Psi(x)\|^2 = \frac{1}{K} \sum_{k=1}^{K} \kappa(x, \alpha_k)^2 \tag{20}$$

Let $X_k = \kappa(x, \alpha_k)^2$. Since the Gaussian kernel is strictly bounded in $[0, 1]$, each X_k is an independent random variable bounded such that $0 \leq X_k \leq 1$. Let $\mu_{self} = \mathbb{E}_{\alpha \sim \mathcal{D}_0}[\kappa(x, \alpha)^2]$ be the expected energy. We invoke Hoeffding's Inequality for bounded independent random variables. For any $\epsilon > 0$

$$\mathbb{P}\left(|Z_K - \mu_{self}| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2K^2\epsilon^2}{\sum_{k=1}^{K}(1-0)^2}\right) \tag{21}$$

:Simplifying the denominator:

$$\mathbb{P}\left(\left|\frac{1}{K}\|\Psi(x)\|^2 - \mu_{self}\right| \geq \epsilon\right) \leq 2\exp(-2K\epsilon^2) \tag{22}$$

Interpretation: The probability that the empirical feature norm deviates from the constant expectation μ_{self} decays exponentially with K. This implies that as K increases (e.g., $K = 7000$), all feature vectors almost surely reside on a stable hyperspherical shell of radius $R \approx \sqrt{K\mu_{self}}$, effectively eliminating magnitude variance caused by few-shot scarcity.

Proof of Strict Linear Separability. We analyze the expected inner product (cross-correlation) between two distinct samples x and x' in the RKHS. Let the empirical inner product be:

$$S_K = \frac{1}{K}\langle\Psi(x), \Psi(x')\rangle = \frac{1}{K}\sum_{k=1}^{K}\kappa(x, \alpha_k)\kappa(x', \alpha_k) \tag{23}$$

By the Law of Large Numbers, as $K \to \infty$, S_K converges to its expectation:

$$\mu_{cross} = \mathbb{E}_{c\sim\mathcal{D}_0}[\kappa(x, \alpha)\kappa(x', \alpha)] \tag{24}$$

To prove separability, we compare μ_{cross} with the self-energy μ_{self} using the Cauchy-Schwarz Inequality for expectations:

$$|\mathbb{E}[f(c)g(c)]| \leq \sqrt{\mathbb{E}[f(c)^2] \cdot \mathbb{E}[g(c)^2]} \tag{25}$$

Let $f(c) = \kappa(x, c)$ and $g(c) = \kappa(x', c)$. Substituting these into the inequality:

$$\mu_{cross} \leq \sqrt{\mu_{self}(x) \cdot \mu_{self}(x')} \tag{26}$$

Strict Inequality Condition: The equality in Cauchy-Schwarz holds if and only if $f(c)$ and $g(c)$ are linearly dependent, i.e., $\kappa(x, c) = \lambda\kappa(x', c)$ for almost all c. Given the local sensitivity of the Gaussian RBF kernel, for any two distinct semantic concepts $x \neq x'$, the functions $\kappa(x, \cdot)$ and $\kappa(x', \cdot)$ are linearly independent. Thus, the inequality is strict:

$$\mu_{cross} < \sqrt{\mu_{self}(x)\mu_{self}(x')} \tag{27}$$

Conclusion:This "Energy Gap" implies that the projection of any sample onto itself (intra-class similarity) is strictly larger than its projection onto a distinct class sample (inter-class similarity). As K increases, the variance of these estimates vanishes (due to Hoeffding's bound), rendering this separation deterministic. Consequently, the Gram matrix is diagonally dominant, ensuring that a linear solver can find an optimal decision boundary.

C Detailed Derivation Process of Global Knowledge-Aware Solution

In this section, we provide detailed mathematical derivation for the Global Knowledge-Aware Solver (GKAS), illustrating that how GKAS analytically updates the prototypes and achieves global optimal weights without retraining, eliminating the risk of catastrophic forgetting.

C.1 Derivation Details

We firstly introduce two fundamental linear algebra concepts utilized in our method: Cholesky Decomposition and the Woodbury Matrix Identity.

Cholesky Decomposition. Cholesky decomposition factorizes a Hermitian positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose, which is useful for efficient numerical solutions. The formula for the Cholesky decomposition is expressed as follows:

$$\mathcal{A} = LL^\top \tag{28}$$

where $\mathcal{A} \in \mathbb{R}^{n\times n}$ is a Hermitian positive-definite matrix, and $L \in \mathbb{R}^{n\times n}$ is a real lower triangular matrix with positive diagonal entries.

Woodbury Matrix Identity. This identity is an important identity used to compute the inverse of a matrix to ensure numerical stability and accuracy in linear algebra and numerical computation. This identity is the cornerstone of our recursive update mechanism. It is formulated as:

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \tag{29}$$

where $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{n\times n}$ are all conformable matrices.

Proposition 3.3 Recursive Update Solver. To endow the prototypes the dynamical capability and address the extreme imbalance in FSCIL, we formulate the optimization problem as Regularized Weighted Least Squares problem. This rebalances the contribution of different classes, preventing the model from being dominated by base classes with abundant samples. This formulation can be formulated as:

$$\mathcal{L}_{\mathcal{W}_t} = \|\Omega_{0:t}^{\frac{1}{2}}(\mathcal{Y}_{0:t} - \Phi_{0:t}\mathcal{W}_t)\|_F^2 + \lambda\|\mathcal{W}_t\|_F^2, \tag{30}$$

where $\Phi_{0:t} \in \mathbb{R}^{N_{total}\times K}$ denotes the accumulated feature matrix of all samples up to session t. $\mathcal{Y}_{0:t}$ is the corresponding one-hot label matrix. $\mathcal{W}_t \in \mathbb{R}^{K\times C_{total}}$ is the projection matrix (classifier weights) to be optimized and $\lambda > 0$ is the regularization coefficient that ensures numerical stability. $\Omega_{0:t} \in \mathbb{R}^{N_{total}\times N_{total}}$ is a diagonal weighted matrix, which gives each sample a weight equal to inversely proportional to its class frequency and formulates as:

$$\Omega_t = \begin{bmatrix} 1/N_0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 1/N_k & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1/N_K \end{bmatrix} \tag{31}$$

Since Eq. (30) is a strictly convex objective function, we can derive the global optimal solution $\mathcal{W}_{t+1}$ at session $t+1$(where $t \geq 0$) by setting the gradient w.r.t. $\mathcal{W}_t$ to zero:

$$\mathcal{W}_{t+1}^* = \left(\Phi_{0:t+1}^\top \Omega_{0:t+1}\Phi_{0:t+1} + \lambda I\right)^{-1} \mu_{0:t+1}^\top \tag{32}$$

where

$$\mu_{0:t+1}^\top = \Phi_{0:t+1}^\top \Omega_{0:t+1}\mathcal{Y}_{0:t+1} \tag{33}$$

Here, $\mathcal{A}_{t+1} = \Phi_{0:t+1}^\top \Omega_{0:t+1}\Phi_{0:t+1} + \lambda I$ denotes the global knowledge gram matrix, μ_{t+1} is mathematically equivalent to the empirical prototypes of all seen classes. By multiplying with $\mathcal{A}_{t+1}^{-1}$, the prototypes explicitly suppress the interference among correlated categories, rectifying the local empirical means into globally optimized decision anchors. Consequently, we term the resulting projection matrix $\mathcal{W}^*$ as Dynamic Analytical Prototypes.

As stated in the problem definition, historical samples $\mathbf{\Phi}_{0:t}$ are unavailable at session $t+1$. We update the global optimal weights solely rely on new samples Φ_{t+1}, and the statistics from the previous session $(\mathcal{P}_t, \mathcal{W}_t)$. The global knowledge matrix can be decomposed into the historical component and the incremental component:

$$\begin{aligned} \mathcal{A}_{t+1} &= \lambda I + \begin{bmatrix}\Phi_{0:t} & \Phi_{t+1}\end{bmatrix}^\top \begin{bmatrix}\Omega_{0:t} & 0 \\ 0 & \Omega_{t+1}\end{bmatrix} \begin{bmatrix}\Phi_{0:t} \\ \Phi_{t+1}\end{bmatrix} \\ &= \lambda I + \Phi_{0:t}^\top \Omega_{0:t}\Phi_{0:t} + \Phi_{t+1}^\top \Omega_{t+1}\Phi_{t+1} \\ &= \mathcal{A}_t + \Phi_{t+1}^\top \Omega_{t+1}\Phi_{t+1} \end{aligned} \tag{34}$$

Furthermore, when taking a positive term λI, the matrix $\mathcal{A}$ is definite positive. So the inverse of $\mathcal{A}$ is also a definite positive matrix, which can be decomposed and expanded by Eq.(28) and Eq.(29) as:

$$\begin{aligned} \mathcal{A}_{t+1}^{-1} &= \mathcal{P}_{t+1}\mathcal{P}_{t+1}^\top \\ &= \mathcal{A}_t^{-1} - \mathcal{A}_t^{-1}\Phi_{t+1}^\top \\ &\quad (\Omega_{t+1}^{-1} + \Phi_{t+1}\mathcal{A}_t^{-1}\Phi_{t+1}^\top)^{-1}\Phi_{t+1}\mathcal{A}_t^{-1} \\ &= \mathcal{P}_t(I - \mathcal{Q}_t^\top(\Omega_{t+1}^{-1} + \mathcal{Q}_t\mathcal{Q}_t^\top)^{-1}\mathcal{Q}_t)\mathcal{P}_t^\top \end{aligned} \tag{35}$$

Backbone	Trainable components	Optimizer	Epochs	Feature LR	Head LR	Weight decay	Min. LR	Label smoothing
VIT-B-IN21K	AdaptFormer adapters only, with FFN dimension 64; pretrained backbone frozen	SGD	40	1×10^{-2}	1×10^{-2}	5×10^{-4}	0	0
ResNet-18	Entire backbone	SGD	40	1×10^{-3}	1×10^{-2}	5×10^{-4}	0	0.1
CLIP-VIT-B	Last two Transformer blocks, final normalization layer, and visual projection	AdamW	20	1×10^{-5}	1×10^{-3}	5×10^{-2}	1×10^{-7}	0.1

Table 4: Base-session training configurations for different backbones. Feature LR denotes the learning rate of the trainable feature-extraction parameters, while Head LR denotes that of the temporary linear classifier.

and

$$\mathcal{Q}_t = \Phi_{t+1}\mathcal{P}_t \tag{36}$$

Similarly, the prototype matrix can be updated incrementally as:

$$\begin{aligned}\mu_{0:t+1}^\top &= \begin{bmatrix}\Phi_{0:t}^\top & \Phi_{t+1}^\top\end{bmatrix}\begin{bmatrix}\Omega_{0:t} & 0\\ 0 & \Omega_{t+1}\end{bmatrix}\begin{bmatrix}\mathcal{Y}_{0:t} & 0\\ 0 & \mathcal{Y}_{t+1}\end{bmatrix}\\ &= \begin{bmatrix}\mathcal{A}_t\mathcal{A}_t^{-1}\Phi_{0:t}^\top\Omega_{0:t}\mathcal{Y}_{0:t} & \Phi_{t+1}^\top\Omega_{t+1}\mathcal{Y}_{t+1}\end{bmatrix}\\ &= \begin{bmatrix}\mathcal{A}_t\mathcal{W}_t & \mu_{t+1}^\top\end{bmatrix}\end{aligned} \tag{37}$$

From Eq.(32),Eq.(33),Eq.(34) and Eq.(37), the global optimal weights $\mathcal{W}_{n+1}$ for session $t+1$ are obtained by:

$$\begin{aligned}\mathcal{W}_{t+1} &= \mathcal{A}_{t+1}^{-1}\left[\mathcal{A}_t\mathcal{W}_t\ \ \mu_{t+1}^\top\right]\\ &= \mathcal{A}_{t+1}^{-1}\left[(\mathcal{A}_{t+1}-\Phi_{t+1}^\top\Omega_{t+1}\Phi_{t+1})\mathcal{W}_{t+1}\ \ \mu_{t+1}^\top\right]\\ &= \left[\mathcal{W}_t-\mathcal{A}_{t+1}^{-1}\Phi_{t+1}^\top\Omega_{t+1}\Phi_{t+1}\mathcal{W}_t\ \ \mathcal{A}_{t+1}^{-1}\mu_{t+1}^\top\right]\end{aligned} \tag{38}$$

Mathematically, the derivation proves that the global knowledge-aware solver calculated recursively is mathematically identical to the solution joint training on entire dataset $\Phi_{0:t+1}$, effectively disentangles the dense distribution and suppresses the noise induced by high dimensionality, ensuring robust separability where heuristic static classifiers fail. This mechanism explicitly transforms the static prototypes into a dynamic analytical system, achieving global optimal solution and guaranteeing theoretical zero-forgetting accumulated knowledge as the classifier level.

Proposition 3.4 Explicit Decorrelation. Quoting the theoretical analysis from RanPAC McDonnell et al. (2023a), the weight solving process using the inverse Gram matrix $\mathcal{A}^{-1} = (\Phi^\top\Phi)^{-1}$ explicitly acts as a decorrelation (whitening) transformation. The Ridge Regression solution is $\mathcal{W}_{t+1} = \mathcal{A}_{t+1}^{-1}\mu_{0:t+1}$. We define a linear whitening transform matrix $D \in \mathbb{R}^{M\times M}$ such that:

$$(D\Phi)(D\Phi)^\top \approx D\mathcal{A}D^\top = I \implies D^\top D = \mathcal{A}^{-1} \tag{39}$$

The classification score s_y for a test sample Φ against a prototype μ_y becomes:

$$s_y = \Phi^\top\mathcal{A}^{-1}\mu_y = \Phi^\top(D^\top D)\mu_y = (D\Phi)^\top(D\mu_y) \tag{40}$$

Since D whitens the data distribution ($D\mathcal{A}D^\top = I$), the transformed prototypes $D\mu_y$ have an identity covariance. This explicitly minimizes the Pearson correlation coefficients between class prototypes, solving the issue of high inter-class correlation observed in raw pre-trained features. This derivation proves that our method is mathematically equivalent to calculating dot products in a whitened space where the covariance matrix is Identity (I). This effectively removes the off-diagonal elements in the correlation matrix, explicitly minimizing the Pearson correlation coefficients between class prototypes.

D Detailed Training Configuration

Datasets Protocols. We evaluate DAP on three complementary FSCIL benchmarks: CIFAR100, CUB200, and ImageNetR. CIFAR100 Krizhevsky et al. (2009) contains 100 generic object categories and is used to evaluate class-incremental learning under relatively low-resolution visual

Dataset	Base classes	Incremental protocol	Sessions
CIFAR100	60	8× 5-way 5-shot	9
CUB200	100	10× 10-way 5-shot	11
ImageNetR	100	10× 10-way 5-shot	11

Table 5: Dataset protocols used in our experiments.

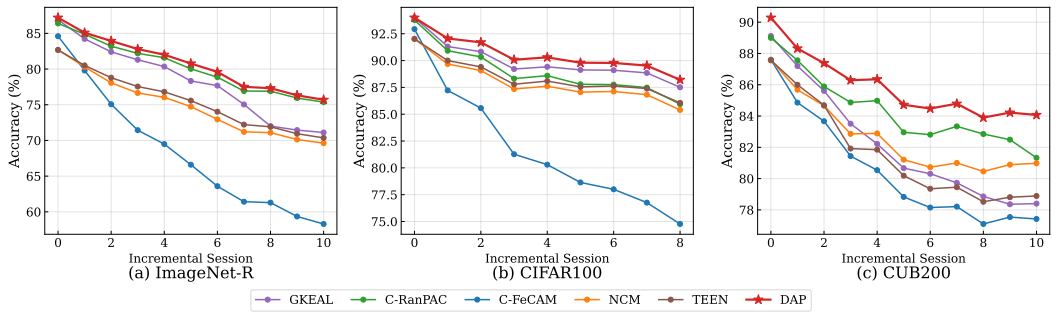

Figure 4: Session-wise accuracy comparison on ImageNetR, CIFAR100, and CUB200 using the fine-tuned VIT-B-IN21K backbone. DAP achieves the highest or tied-highest accuracy throughout the incremental sessions, demonstrating consistent effectiveness across different visual domains.

representations. Following the standard protocol, 60 classes are used in the base session, while the remaining 40 classes are introduced through eight 5-way 5-shot incremental sessions. CUB200 Wah et al. (2011) is a fine-grained bird recognition dataset containing 200 species with subtle inter-class differences. We use 100 base classes and introduce the remaining 100 classes through ten 10-way 5-shot sessions. ImageNetR Hendrycks et al. (2021) contains diverse artistic renditions of 200 ImageNet categories and provides a more challenging evaluation under substantial appearance variation. We adopt a stratified 80/20 training–testing split, with 100 base classes followed by ten 10-way 5-shot incremental sessions. Together, these datasets evaluate DAP across generic recognition, fine-grained classification, and distribution-shifted visual domains.

Backbone Adaptation. For each backbone, adaptation is performed exclusively on the base session using a backbone-specific training strategy shared across all datasets. After base training, the feature extractor is frozen, and identical extracted features are provided to NCM Guerriero et al. (2018), TEEN Wang et al. (2023), GKEAL Zhuang et al. (2023), and the higher-order calibration variants C-FeCAM and C-RanPAC Goswami et al. (2024) for controlled comparison. A newly initialized linear classifier is used only to supervise base-session adaptation and is discarded before incremental learning. Detailed configurations are reported in Table 4.

Optimization and Reproducibility. All models are optimized using cross-entropy loss and a cosine annealing schedule, where $T_{\max}$ equals the total number of training epochs and $\eta_{\min}$ follows Table 4. For SGD, we use a momentum of 0.9 with Nesterov acceleration. The training and feature-extraction batch sizes are 64 and 128, respectively, with eight data-loading workers. All images are resized to 224×224. Mixed-precision training is enabled, and the checkpoint from the final epoch is used without validation-based model selection.

We use a fixed random seed of 1993 and enable deterministic cuDNN execution by setting `torch.backends.cudnn.deterministic=True` and `torch.backends.cudnn.benchmark=False`. Training augmentation consists of a bicubic random resized crop with scale $(0.5, 1.0)$ and aspect-ratio range $(0.75, 1.3333)$, followed by random horizontal flipping. Backbone-specific normalization is applied: mean and standard deviation $(0.5, 0.5, 0.5)$ for ViT-B/16, the standard ImageNet statistics for ResNet-18, and the OpenAI CLIP statistics for CLIP-VIT-B.

Methods	Backbone	Accuracy (%) in each session									Average Acc	PD
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**		
NCM	ResNet-18	78.72	75.31	72.04	68.37	67.00	64.67	63.90	62.48	60.16	68.07	18.56
TEEN	ResNet-18	78.72	75.52	72.51	69.03	67.90	65.84	65.18	63.78	61.44	68.88	17.28
GKEAL	ResNet-18	83.53	77.88	72.96	67.79	63.39	59.46	56.67	54.01	51.18	65.21	32.35
C-FeCAM	ResNet-18	80.35	77.03	74.01	70.23	68.84	66.59	65.97	64.56	61.17	69.86	19.18
C-RanPAC	ResNet-18	82.73	78.63	75.14	70.97	69.14	66.32	64.99	63.49	61.15	70.28	21.58
DAP	ResNet-18	**83.60**	**80.20**	**77.03**	**73.21**	**71.97**	**69.60**	**68.66**	**67.17**	**64.82**	**72.92**	**18.78**
NCM	VIT-B-IN21K	92.02	89.68	89.07	87.35	87.61	87.06	87.12	86.82	85.40	88.01	6.62
TEEN	VIT-B-IN21K	92.02	90.00	89.41	87.79	88.09	87.55	87.63	87.39	86.06	88.44	5.96
C-FeCAM	VIT-B-IN21K	92.93	87.22	85.57	81.27	80.30	78.64	78.00	76.76	74.77	81.72	18.16
C-RanPAC	VIT-B-IN21K	93.78	90.92	90.34	88.32	88.59	87.79	87.76	87.48	85.94	88.99	7.84
GKEAL	VIT-B-IN21K	**94.00**	91.31	90.83	89.21	89.42	89.13	89.10	88.85	87.51	89.93	6.49
DAP	VIT-B-IN21K	**94.00**	**92.06**	**91.70**	**90.08**	**90.30**	**89.80**	**89.78**	**89.53**	**88.21**	**90.61**	**5.79**
NCM	CLIP-VIT-B	85.50	82.00	79.26	75.96	74.10	72.35	71.42	69.85	67.92	75.37	17.58
TEEN	CLIP-VIT-B	85.50	82.12	79.46	76.20	74.56	72.85	71.97	70.46	68.57	75.74	16.93
GKEAL	CLIP-VIT-B	86.87	82.77	80.39	77.19	75.01	73.71	72.67	71.40	69.85	76.65	17.02
C-FeCAM	CLIP-VIT-B	85.62	82.28	79.69	76.51	74.85	73.16	72.37	70.80	68.94	76.02	16.68
C-RanPAC	CLIP-VIT-B	86.78	82.02	78.99	75.51	73.29	71.07	70.10	68.64	66.60	74.78	20.18
DAP	CLIP-VIT-B	**87.02**	**84.15**	**81.74**	**78.47**	**77.18**	**75.60**	**75.00**	**73.59**	**71.79**	**78.28**	**15.23**
SEC-Prompt	VIT-B-IN1K	91.23	89.78	89.36	87.61	87.59	87.58	87.74	87.28	86.12	88.25	5.11
SEC-Prompt+DAP	VIT-B-IN1K	**93.08**	**92.03**	**91.87**	**90.64**	**90.70**	**90.69**	**90.63**	**90.16**	**89.38**	**91.02**	**3.70**
CASP	VIT-B-IN1K	93.07	91.02	90.27	88.69	88.74	88.45	88.54	88.18	87.04	89.33	6.03
CASP+DAP	VIT-B-IN1K	**94.23**	**92.28**	**91.56**	**89.97**	**90.26**	**89.82**	**89.89**	**89.58**	**88.43**	**90.67**	**5.80**

Table 6: FSCIL performance comparison on CIFAR100.

E Additional Comparison Results

Evaluation Protocol. We provide the complete session-wise results on CIFAR100 and ImageNetR to complement the CUB200 results reported in the main paper. Within each backbone group, all post-backbone methods use the same frozen feature extractor, class order, five-shot support sets, and random seeds. Therefore, their performance differences originate from feature calibration, prototype construction, and classifier optimization rather than from differences in representation learning. As in the main paper, all results are averaged over three random seeds. We additionally apply DAP to SEC-Prompt and CASP to examine whether the proposed post-backbone optimization remains effective when combined with existing FSCIL representation-learning frameworks.

Figure 4 presents the session-wise accuracy curves on ImageNetR, CIFAR100, and CUB200 using the same fine-tuned VIT-B-IN21K backbone. DAP maintains the highest or tied-highest accuracy across all incremental sessions on the three datasets. Moreover, its advantage is consistently preserved as novel classes accumulate, indicating that the improvement is not dominated by the base-session performance but is sustained throughout the incremental process. These results further demonstrate the effectiveness of DAP in balancing novel-class acquisition and accumulated classification performance.

E.1 Results on CIFAR100

Table 6 reports the complete session-wise results on CIFAR100. DAP achieves the highest absolute accuracy across all incremental sessions under the three backbone families, demonstrating that its effectiveness is not tied to a particular representation architecture.

With ResNet-18, DAP obtains an AVG of 72.92%, outperforming the strongest competing post-backbone method, C-RanPAC, by 2.64 percentage points. The advantage becomes increasingly evident as additional classes are introduced: DAP reaches 64.82% in the final session, compared with 61.44% for TEEN and 61.15% for C-RanPAC. Although TEEN exhibits a slightly lower PD, DAP maintains substantially higher absolute accuracy throughout the incremental process. This observation indicates that PD alone does not fully characterize FSCIL performance, because a method may obtain a small drop from a relatively weak initial classifier. DAP instead maintains a stronger bal-

Methods	Backbone	Accuracy (%) in each session											Average Acc	PD
		0	1	2	3	4	5	6	7	8	9	10		
NCM	ResNet-18	68.90	65.27	61.05	58.40	56.14	53.20	50.83	48.70	48.10	46.74	45.89	54.84	23.01
TEEN	ResNet-18	68.90	64.80	61.69	59.02	57.10	54.11	51.83	49.78	49.05	47.71	47.04	55.55	21.86
GKEAL	ResNet-18	70.89	67.21	63.12	61.28	58.40	55.51	53.14	52.31	50.54	49.12	48.41	57.27	22.48
C-FeCAM	ResNet-18	68.97	65.75	62.14	59.65	57.42	54.58	52.13	50.14	49.63	48.41	47.55	56.03	**21.42**
C-RanPAC	ResNet-18	70.21	66.52	62.70	60.18	57.89	54.51	52.57	50.54	49.97	48.57	47.65	56.48	22.56
DAP	ResNet-18	**71.38**	**68.39**	**64.52**	**62.05**	**60.01**	**56.71**	**54.26**	**52.08**	**51.46**	**50.00**	**49.29**	**58.20**	22.08
NCM	VIT-B-IN21K	82.68	80.30	78.07	76.65	76.04	74.74	72.95	71.21	71.09	70.13	69.62	74.86	13.06
TEEN	VIT-B-IN21K	82.68	80.51	78.79	77.55	76.81	75.58	74.03	72.22	71.91	70.94	70.36	75.58	12.32
GKEAL	VIT-B-IN21K	86.91	84.23	82.41	81.30	80.35	78.33	77.68	75.04	72.01	71.45	71.11	78.26	15.80
C-FeCAM	VIT-B-IN21K	84.60	79.80	75.07	71.44	69.49	66.60	63.59	61.42	61.29	59.35	58.30	68.27	26.30
C-RanPAC	VIT-B-IN21K	86.39	84.88	83.20	82.20	81.59	80.03	78.86	76.92	76.90	75.94	75.37	80.21	**11.02**
DAP	VIT-B-IN21K	**87.20**	**85.09**	**83.94**	**82.80**	**82.01**	**80.78**	**79.60**	**77.51**	**77.32**	**76.32**	**75.70**	**80.75**	11.50
NCM	CLIP-VIT-B	82.61	79.71	76.06	73.94	71.80	69.95	67.94	65.68	65.26	64.40	63.38	70.98	19.23
TEEN	CLIP-VIT-B	82.61	79.80	76.78	74.82	72.66	70.68	68.48	65.91	65.40	64.54	63.55	71.38	19.06
GKEAL	CLIP-VIT-B	81.14	80.52	77.29	76.40	75.20	73.18	71.35	70.34	67.89	65.27	64.32	72.99	16.82
C-FeCAM	CLIP-VIT-B	82.58	80.04	76.73	74.89	72.99	71.26	69.18	66.89	66.47	65.70	64.75	71.95	17.83
C-RanPAC	CLIP-VIT-B	83.72	80.60	77.34	75.47	73.71	71.86	69.69	67.50	67.09	66.54	65.93	72.68	17.79
DAP	CLIP-VIT-B	**84.01**	**81.34**	**78.65**	**77.07**	**75.18**	**73.22**	**71.09**	**69.18**	**68.80**	**68.06**	**67.18**	**73.98**	**16.83**
SEC-Prompt	VIT-B-IN1K	84.30	81.91	81.36	79.20	78.63	77.90	76.55	76.16	76.08	74.99	74.25	78.30	10.05
SEC-Prompt+DAP	VIT-B-IN1K	**86.15**	**84.12**	**83.99**	**82.21**	**81.75**	**81.08**	**79.91**	**79.37**	**79.05**	**78.35**	**77.68**	**81.24**	**8.46**
CASP	VIT-B-IN1K	82.20	79.17	78.65	76.25	75.45	74.61	73.44	73.09	72.70	71.93	70.95	75.31	11.25
CASP+DAP	VIT-B-IN1K	**85.32**	**82.54**	**82.27**	**80.12**	**79.34**	**78.48**	**77.25**	**76.56**	**76.03**	**75.09**	**74.08**	**78.83**	**11.24**

Table 7: ImageNetR class-incremental comparison.

ance between initial representation quality, novel-class acquisition, and accumulated classification performance.

The advantage of DAP is also preserved with stronger pretrained representations. Using VIT-B-IN21K, DAP achieves an AVG of 90.50%, improving over GKEAL by 0.57 percentage points. It ties GKEAL in the base session and achieves the highest accuracy in every subsequent session, while also reducing PD from 6.49% to 5.79%. These results show that class-balanced analytical optimization remains beneficial even when the frozen representation already provides strong generalization.

Under CLIP-ViT-B, DAP obtains an AVG of 78.28%, exceeding GKEAL by 1.63 percentage points. DAP also achieves the highest final-session accuracy of 71.79% and the lowest PD of 15.23% among the compared methods. The consistent improvements across ResNet, supervised ViT, and vision–language pretrained representations demonstrate that DAP operates effectively as a general post-backbone optimization framework.

We further integrate DAP into SEC-Prompt and CASP. DAP improves the AVG of SEC-Prompt from 88.25% to 91.02% and reduces its PD from 5.11% to 3.70%. When combined with CASP, DAP improves AVG from 89.33% to 90.67% and reduces PD from 6.03% to 5.80%. These improvements indicate that DAP is complementary to representation-learning approaches: it can further refine their frozen features and construct more coordinated decision boundaries without modifying their original training objectives.

E.2 Results on ImageNetR

Table 7 presents the complete results on ImageNetR. Compared with CIFAR100 and CUB200, ImageNetR contains substantial appearance variations and artistic renditions, making it a more challenging test of whether the calibrated statistics and global classifier remain effective under distribution-shifted visual representations. DAP achieves the highest absolute accuracy in every session for all three backbones.

With ResNet-18, DAP achieves an AVG of 58.20%, improving over GKEAL by 0.93 percentage points and over NCM by 3.36 points. Its final-session accuracy reaches 49.29%, compared with 48.41% for GKEAL and 47.65% for C-RanPAC. C-FeCAM obtains a slightly smaller PD, but DAP maintains higher accuracy from the base session to the final session. This result again highlights that the lowest PD does not necessarily correspond to the strongest accumulated classifier; DAP achieves a more favorable trade-off by learning new classes while retaining high absolute performance over all observed classes.

Using VIT-B-IN21K, DAP obtains an AVG of 80.75%, exceeding C-RanPAC by 0.54 percentage points. Although C-RanPAC has a marginally lower PD, DAP achieves the highest accuracy in every session and improves the final-session accuracy from 75.37% to 75.70%. The result suggests that the calibrated distributions and class-balanced analytical objective provide complementary gains beyond the high-quality frozen ViT representation.

With CLIP-ViT-B, DAP achieves an AVG of 73.98%, improving over GKEAL by 0.99 percentage points. Its final-session accuracy of 67.18% is also 1.25 points higher than that of C-RanPAC and 2.86 points higher than that of GKEAL. The PD of DAP is nearly identical to that of GKEAL, while DAP maintains consistently higher absolute accuracy. Thus, the improvement does not arise merely from changing the initial base-session accuracy, but is sustained throughout the incremental process.

DAP also improves the two prompt-based FSCIL frameworks on ImageNetR. When applied to SEC-Prompt, DAP increases AVG from 78.30% to 81.24% and reduces PD from 10.05% to 8.47%. When applied to CASP, DAP improves AVG from 75.31% to 78.83%, while maintaining a comparable PD (11.25% versus 11.24%). In the latter case, the nearly unchanged PD should be interpreted together with the substantial increase in absolute accuracy: DAP improves both the initial classifier and the subsequent session-wise performance without increasing the relative degradation.

E.3 Resuts On CUB200

Table 8 presents the complete results on CUB200. As a fine-grained bird recognition benchmark, CUB200 tests whether DAP can distinguish closely related categories with limited novel-class observations. DAP achieves the highest AVG and final-session accuracy within each of the three backbone groups.

With ResNet-18, DAP achieves an AVG of 65.29%, improving over GKEAL by 0.27 percentage points and over NCM by 2.11 points. Its final-session accuracy reaches 58.25%, compared with 57.68% for GKEAL and 57.42% for C-RanPAC. Although GKEAL performs better in session 6, DAP obtains the highest or tied-highest accuracy in the remaining sessions and the lowest reported PD of 18.85%. These results show a modest overall improvement over the strongest competing method with the CNN representation.

Using VIT-B-IN21K, DAP obtains an AVG of 85.89%, exceeding C-RanPAC by 1.52 percentage points. DAP achieves the highest accuracy in every session, improves the final-session accuracy from 81.33% to 84.07%, and reduces PD from 7.67% to 6.22%. Its advantage over C-RanPAC increases from 1.29 points in the base session to 2.74 points in the final session, indicating that the improvement extends beyond the initial classifier and remains effective as novel classes accumulate.

With CLIP-ViT-B, DAP achieves an AVG of 74.52%, improving over C-RanPAC by 0.88 percentage points. Its final-session accuracy reaches 69.30%, compared with 68.55% for C-RanPAC and 68.05% for GKEAL. Although C-FeCAM obtains a smaller PD (14.04% versus 14.33%), DAP maintains the highest absolute accuracy in every session. This result again highlights that a lower PD does not necessarily imply stronger overall classification performance; the drop should be interpreted together with the accuracy maintained throughout incremental learning.

DAP also improves the two prompt-based FSCIL frameworks on CUB200. When applied to CASP, DAP increases AVG from 83.92% to 86.12% and improves the final-session accuracy from 82.91% to 84.79%. When applied to SEC-Prompt, DAP raises AVG from 82.45% to 86.22% and final-session accuracy from 81.42% to 85.16%. PD remains comparable, decreasing slightly from 5.49% to 5.43% for CASP and from 5.17% to 5.11% for SEC-Prompt. These results support DAP as a complementary post-backbone optimization module that improves session-wise classification performance without increasing the base-to-final accuracy drop.

F Extended Ablation Studies

Experimental Setup. We conduct exhaustive component ablations on CIFAR100, CUB200, and ImageNetR using the fine-tuned VIT-B-IN21K backbone. We evaluate all 2^3 combinations of Distribution Calibration (DC), Kernel Mapping (KM), and the Global Knowledge-Aware Solver (GKAS). All variants use the same frozen representations, class orders, few-shot support sets, and hyperpa-

Methods	Backbone	Accuracy (%) in each session											Average Acc	PD
		0	1	2	3	4	5	6	7	8	9	10		
NCM	ResNet-18	74.65	71.40	68.45	65.37	63.65	61.49	60.62	59.09	57.40	57.21	55.66	63.18	18.99
TEEN	ResNet-18	74.65	71.27	68.51	64.65	62.88	60.77	59.80	58.07	56.09	55.95	55.42	62.55	19.23
GKEAL	ResNet-18	77.09	73.72	71.05	67.25	65.11	63.03	62.76	60.53	58.71	58.31	57.68	65.02	19.41
C-RanPAC	ResNet-18	77.03	73.37	70.23	66.52	65.44	63.15	61.57	60.62	58.47	58.10	57.42	64.72	19.61
C-FeCAM	ResNet-18	75.98	72.96	70.32	67.17	65.71	63.59	61.98	60.58	58.62	58.59	57.06	64.78	18.92
DAP	ResNet-18	**77.09**	**73.94**	**71.05**	**67.36**	**65.81**	**63.73**	**62.29**	**60.84**	**58.99**	**58.83**	**58.25**	**65.29**	**18.85**
NCM	VIT-B-IN21K	87.57	85.68	84.66	82.86	82.89	81.21	80.73	81.00	80.46	80.89	80.98	82.63	6.59
TEEN	VIT-B-IN21K	87.57	86.00	84.69	81.92	81.85	80.19	79.35	79.45	78.52	78.81	78.89	81.57	8.68
GKEAL	VIT-B-IN21K	89.12	87.20	85.61	83.51	82.22	80.67	80.31	79.73	78.86	78.36	78.40	82.18	10.72
C-RanPAC	VIT-B-IN21K	89.00	87.56	85.89	84.87	84.98	82.96	82.81	83.34	82.85	82.49	81.33	84.37	7.67
C-FeCAM	VIT-B-IN21K	87.60	84.86	83.67	81.44	80.54	78.83	78.15	78.21	77.10	77.54	77.42	80.49	10.18
DAP	VIT-B-IN21K	**90.29**	**88.32**	**87.38**	**86.29**	**86.35**	**84.72**	**84.48**	**84.79**	**83.90**	**84.22**	**84.07**	**85.89**	**6.22**
NCM	CLIP-VIT-B	82.26	79.86	77.73	74.36	73.39	71.61	70.20	69.23	67.67	67.86	67.62	72.89	14.64
TEEN	CLIP-VIT-B	82.26	79.38	77.06	73.31	72.24	70.41	68.99	67.77	65.98	66.11	65.76	71.75	16.50
GKEAL	CLIP-VIT-B	82.44	79.73	77.96	74.95	74.16	72.17	71.00	70.07	68.38	68.71	68.05	73.42	14.39
C-RanPAC	CLIP-VIT-B	82.72	80.40	78.16	74.84	74.21	72.21	70.91	70.35	68.59	69.08	68.55	73.64	14.17
C-FeCAM	CLIP-VIT-B	82.23	79.92	78.05	74.79	73.88	72.19	70.85	69.95	68.27	68.42	68.19	73.34	**14.04**
DAP	CLIP-VIT-B	**83.62**	**80.75**	**79.15**	**75.99**	**75.15**	**73.16**	**72.24**	**71.11**	**69.46**	**69.77**	**69.30**	**74.52**	14.33
CASP	VIT-B-IN1K	88.41	86.38	85.25	84.01	84.11	82.39	82.19	82.63	82.17	82.65	82.91	83.92	5.49
CASP+DAP	VIT-B-IN1K	**90.22**	**88.36**	**87.26**	**86.24**	**86.49**	**84.91**	**84.74**	**85.16**	**84.52**	**84.62**	**84.79**	**86.12**	**5.43**
SEC-Prompt	VIT-B-IN1K	86.59	85.12	83.99	82.66	82.32	81.25	80.90	81.38	80.50	80.81	81.42	82.45	5.17
SEC-Prompt+DAP	VIT-B-IN1K	**90.26**	**89.06**	**87.37**	**86.45**	**85.99**	**85.02**	**84.83**	**84.94**	**84.67**	**84.70**	**85.16**	**86.22**	**5.11**

Table 8: FSCIL performance comparison on CUB200. PD denotes the performance drop between the last and first session. **Bold Value** indicates the best classification results.

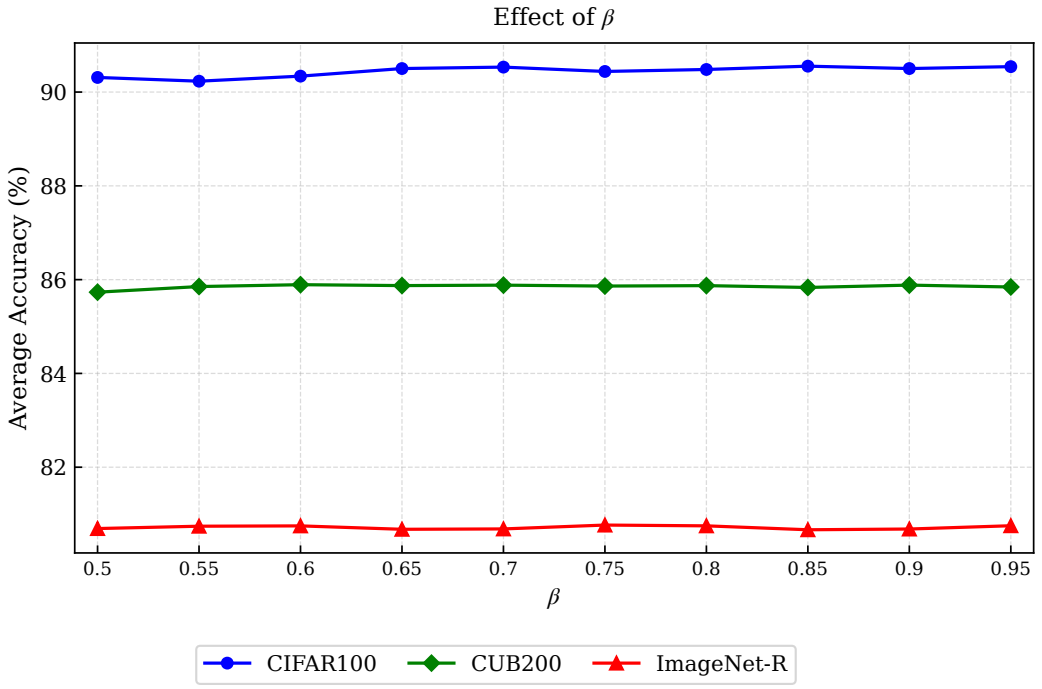

Figure 5: Sensitivity analysis of the covariance scaling coefficient β on CIFAR100, CUB200, and ImageNetR. DAP maintains stable average accuracy over a broad range of β, demonstrating limited sensitivity to the covariance scale.

rameter selection protocol. These experiments separately examine the contribution of each component and, more importantly, their functional dependencies within the complete framework.

Sensitivity to β We further provide the parameter analysis of parameter β in Figure5. The covariance scaling coefficient β controls the overall magnitude of the calibrated covariance, thereby

DC	KM	GKAS	Accuracy (%) in each session									AVG	PD
			0	1	2	3	4	5	6	7	8		
			91.88	89.71	89.06	87.48	87.74	87.19	87.28	86.99	85.55	88.10	6.33
✓			91.88	90.17	89.61	88.21	88.51	87.99	88.06	87.79	86.59	88.76	**5.29**
	✓		86.92	84.12	82.27	79.97	79.80	79.14	78.80	78.25	76.85	80.68	10.07
		✓	93.38	90.71	90.04	88.16	88.20	87.45	87.58	87.26	85.93	88.75	7.45
✓		✓	93.38	91.26	90.84	89.40	89.45	88.96	89.16	88.76	87.44	89.85	5.94
✓	✓		86.92	84.31	81.94	79.44	78.85	77.94	77.42	76.79	75.50	79.90	11.42
	✓	✓	93.97	91.22	90.69	88.92	89.12	88.78	88.64	88.35	86.79	89.61	7.18
✓	✓	✓	**94.00**	**92.06**	**91.70**	**90.08**	**90.30**	**89.80**	**89.78**	**89.53**	**88.21**	**90.61**	**5.79**

Table 9: Ablation results (%) on CIFAR100 using the fine-tuned VIT-B-IN21K backbone.

DC	KM	GKAS	Accuracy (%) in each session											AVG	PD
			0	1	2	3	4	5	6	7	8	9	10		
			87.64	85.78	84.72	82.99	83.12	81.44	80.95	81.19	80.66	81.02	81.10	82.78	6.54
✓			87.64	86.03	85.01	82.99	83.02	81.32	80.82	81.10	80.46	80.91	80.96	82.75	6.68
	✓		80.20	78.27	76.82	74.46	74.30	72.38	71.37	70.33	69.04	69.20	69.00	73.22	11.20
		✓	88.48	86.35	85.71	84.17	84.01	82.11	81.73	82.35	81.33	81.67	81.57	83.59	6.91
✓		✓	88.48	86.61	85.69	84.01	84.16	82.34	82.01	82.24	81.02	81.62	81.33	83.59	7.15
✓	✓		80.20	77.95	76.53	73.85	73.46	71.54	70.48	69.44	68.02	68.11	67.92	72.50	12.29
	✓	✓	90.29	87.94	87.08	86.05	86.00	84.28	84.05	84.22	83.38	83.82	83.72	85.53	6.57
✓	✓	✓	90.29	88.32	87.38	86.29	86.35	84.72	84.48	84.79	83.90	84.22	84.07	**85.89**	**6.22**

Table 10: Ablation results (%) on CUB200 using the fine-tuned VIT-B-IN21K backbone.

determining the dispersion of the synthesized novel-class features. As shown in Figure 5, the average accuracy remains nearly unchanged when β varies from 0.5 to 0.95 across CIFAR100, CUB200, and ImageNetR. No dataset exhibits a pronounced monotonic trend or a sharply localized optimum, indicating that DAP is insensitive to moderate variations in the covariance scale. This robustness suggests that the transferred covariance structure, rather than a precisely tuned scaling coefficient, provides the main benefit of distribution calibration.

F.1 Ablation Results on CIFAR100

Table 9 presents the complete ablation results on CIFAR100. The NCM baseline obtains an AVG of 88.10%. Applying DC alone improves AVG to 88.76% and reduces PD from 6.33% to 5.29%, showing that base-class priors provide more reliable estimates of novel-class statistics. GKAS alone achieves a similar AVG of 88.75%, indicating that global analytical optimization can improve the frozen representation even without explicit distribution calibration.

In contrast, KM alone reduces AVG to 80.68%. This result does not indicate that the kernel space is intrinsically ineffective; rather, the nonlinear mapping changes the feature geometry and introduces high-dimensional correlated coordinates that static NCM cannot properly exploit. This interpretation is further supported by the combination of KM and GKAS, which raises AVG to 89.61%. GKAS is therefore essential for translating the mapped features into globally coordinated decision weights.

Combining DC and GKAS achieves 89.85% AVG, whereas the complete framework reaches the highest AVG of 90.50%. The full model obtains the highest or tied-highest accuracy in every session and improves the baseline by 2.40 percentage points. Although DC alone yields a slightly smaller PD, the complete model maintains substantially higher absolute accuracy throughout the incremental process. These results demonstrate that DC, KM, and GKAS address complementary limitations: DC improves statistical reliability, KM constructs a shared nonlinear geometry, and GKAS converts this geometry into class-balanced global decisions.

DC	KM	GKAS	Accuracy (%) in each session											AVG	PD
			0	1	2	3	4	5	6	7	8	9	10		
			82.29	80.21	78.20	76.95	76.32	75.15	73.49	71.65	71.65	70.89	70.44	75.20	11.85
✓			82.29	80.10	78.71	77.67	77.11	75.99	74.49	72.75	72.60	71.82	71.47	75.91	10.81
	✓		71.80	68.51	64.23	62.43	61.38	59.65	57.34	55.24	55.01	53.71	53.07	60.21	18.73
		✓	84.60	82.50	80.90	79.42	78.91	77.99	76.60	74.63	74.80	73.85	73.20	77.95	11.40
✓		✓	84.60	82.74	81.70	80.48	79.82	78.89	77.84	75.97	75.89	75.04	74.47	78.86	**10.13**
✓	✓		71.80	68.21	64.79	62.55	61.29	59.35	56.92	54.63	54.53	53.36	52.58	60.00	19.22
	✓	✓	87.20	84.97	83.14	81.98	81.26	79.94	78.68	76.80	76.65	75.46	74.76	80.08	12.44
✓	✓	✓	87.20	85.09	83.94	82.80	82.01	80.78	79.60	77.51	77.32	76.32	75.70	**80.75**	11.50

Table 11: Ablation results (%) on ImageNetR using the fine-tuned VIT-B-IN21K backbone.

F.2 Ablation Results on CUB200

Table 10 reports the complete results on the fine-grained CUB200 benchmark. The NCM baseline obtains an AVG of 82.78%. DC alone produces a comparable AVG of 82.75%, suggesting that statistical calibration by itself is insufficient to resolve the strong inter-class similarity among fine-grained bird categories. GKAS alone improves AVG to 83.59%, confirming the importance of globally coordinating the decision boundaries among closely related classes.

As on CIFAR100, KM cannot be effectively used with static NCM: KM alone and DC+KM obtain only 73.22% and 72.50% AVG, respectively. However, combining KM with GKAS substantially improves AVG to 85.53%, exceeding the baseline by 2.75 points. This contrast shows that the kernel mapping is designed to provide a shared high-dimensional reference geometry, while GKAS is required to suppress correlated directions and construct effective decision weights in that space.

The complete model further improves AVG to 85.89% and reduces PD from 6.57% for KM+GKAS to 6.22%. Although the standalone effect of DC is limited on this fine-grained dataset, its contribution becomes clear when it is coupled with KM and GKAS: calibrated novel-class statistics provide a more reliable basis for constructing the shared geometry and updating the global classifier. The full model consequently achieves the highest or tied-highest accuracy across all sessions.

F.3 Ablation Results on ImageNetR

Table 11 evaluates the component interactions under the substantial appearance variations of ImageNetR. The NCM baseline obtains an AVG of 75.20%. DC alone improves AVG to 75.91% and reduces PD from 11.85% to 10.82%, showing that prior-guided calibration is beneficial when few-shot observations exhibit substantial visual variation. GKAS alone provides a larger improvement, reaching 77.95% AVG, while combining DC and GKAS further increases AVG to 78.86%.

KM alone again performs poorly with a static classifier, obtaining 60.22% AVG. However, its combination with GKAS reaches 80.08%, an improvement of 4.88 points over the NCM baseline. This substantial difference demonstrates that the effectiveness of the kernel mapping depends on global analytical optimization: KM provides a more expressive reference geometry, whereas GKAS is responsible for extracting discriminative directions from this high-dimensional space.

The complete framework obtains the best AVG of 80.75%, improving the baseline by 5.55 percentage points. Adding DC to KM+GKAS increases AVG by a further 0.67 points and reduces PD from 12.44% to 11.50%. The improvement indicates that distribution calibration becomes particularly valuable under appearance shifts, where the empirical statistics estimated from five observations may be less representative of the underlying novel-class distribution.

F.4 Cross-Dataset Component Analysis

The exhaustive ablations reveal a consistent dependency among the three components. First, DC directly improves the reliability of few-shot statistics on CIFAR100 and ImageNetR, while its standalone effect is less pronounced on the fine-grained CUB200 dataset. Its benefit is nevertheless consistently observed when combined with the complete pipeline.

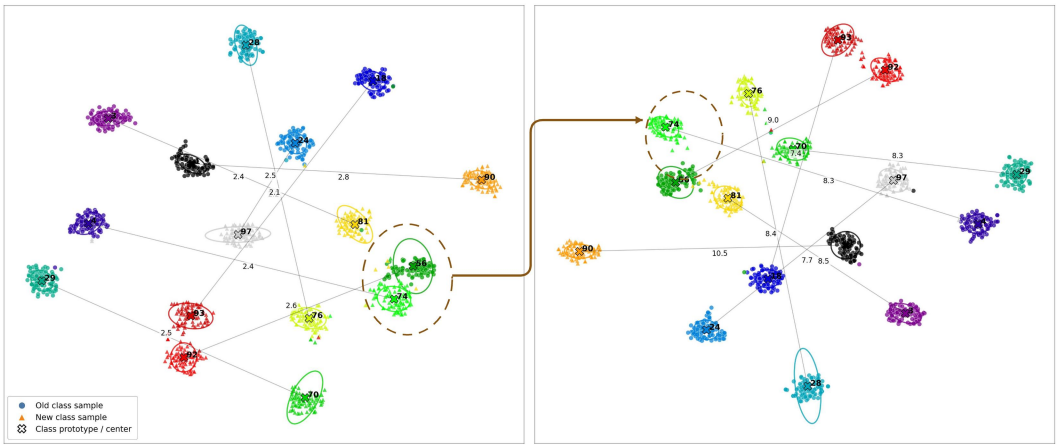

Figure 6: t-SNE visualization before and after feature-space refinement. Numbers on the connecting lines denote distances between selected class centers. The refined space exhibits more compact clusters and larger inter-class margins.

Method	New Acc.	All Acc.	Improved	Negative
5-shot baseline	80.35%	88.10%	0	0
Similar Statistic	81.40%	88.76%	32	8
All-softmax Statistic	79.75%	87.66%	26	10
Random Statistic	78.74%	85.78%	18	22
Dissimilar Statistic	75.84%	85.52%	16	24

Table 12: Failure-case analysis on CIFAR100. Improved/Negative denote the numbers of novel classes whose accuracy increases/decreases after calibration.

Second, KM is not intended to function as an independent classifier. When followed by static NCM, its high-dimensional representation contains correlated and potentially redundant coordinates that cannot be effectively handled by a local prototype rule. This explains the performance degradation of the KM-only and DC+KM variants. Once combined with GKAS, however, KM consistently provides large gains, improving the baseline AVG by 1.51, 2.75, and 4.88 percentage points on CIFAR100, CUB200, and ImageNetR, respectively.

Third, GKAS is the mechanism that converts calibrated and mapped representations into globally coordinated, class-balanced decisions. Its standalone gains show the value of global analytical optimization, while its stronger performance with KM demonstrates the importance of the shared kernel geometry. Finally, the complete model obtains the highest AVG on all three datasets, improving the corresponding baseline by 2.40 points on CIFAR100, 3.11 points on CUB200, and 5.55 points on ImageNetR.

These observations support the sequential design of DAP: DC first improves the statistical reliability of the five-shot novel classes, KM then places historical and novel classes in a shared nonlinear geometry, and GKAS performs class-balanced global updates in that space. Removing any stage breaks at least one connection between statistical reliability, geometric expressiveness, and global discrimination.

Visualization. We employ t-SNE Van der Maaten & Hinton (2008) to visualize old- and novel-class features before and after refinement. The plots include test samples and their corresponding class prototypes, while the annotated values indicate distances between selected prototype pairs. After refinement, both old and novel classes form more compact clusters with larger inter-class distances, demonstrating improved feature separability and prototype consistency.

Effectiveness and Failure Cases of DC. Table 12 examines how the choice of transferred base-class statistics affects distribution calibration. Using statistics from semantically similar base classes

yields the best performance, improving novel-class accuracy from 80.35% to 81.40% and overall accuracy from 88.10% to 88.76%. It also benefits the largest number of novel classes, indicating that relevant base-class priors can effectively reduce the estimation error of five-shot statistics. In contrast, aggregating all base classes or transferring random and dissimilar statistics introduces irrelevant distributional information and leads to increasingly severe negative transfer. These results confirm that DC is effective when the transferred statistics are sufficiently related to the novel class, while also revealing its main limitation under inaccurate prior matching.